Multiclass Classifiers Based on Dimension Reduction with Generalized LDA

Hyunsoo Kim [a] Barry L. Drake [a] Haesun Park [a]

[a] *College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA*

Abstract

Linear discriminant analysis (LDA) has been widely used for dimension reduction of data sets with multiple classes. The LDA has been recently extended to various generalized LDA methods which are applicable regardless of the relative sizes between the data dimension and the number of data items. In this paper, we propose several multiclass classifiers based on generalized LDA algorithms, taking advantage of the dimension reducing transformation matrix without requiring additional training or any parameter optimization. A marginal linear discriminant classifier, a Bayesian linear discriminant classifier, and a one-dimensional Bayesian linear discriminant classifier are introduced for multiclass classification. Our experimental results illustrate that these classifiers produce higher ten-fold cross validation accuracy than kNN and centroid based classification in the reduced dimensional space providing efficient general multiclass classifiers.

Key words: Multiclass Classifier, Generalized Linear Discriminant Analysis (LDA), Generalized Singular Value Decomposition (GSVD), Dimension Reduction, Undersampled Problems

Email addresses: hskim@cc.gatech.edu (Hyunsoo Kim), bldrake@cc.gatech.edu (Barry L. Drake), hpark@cc.gatech.edu (Haesun Park).

1 Introduction

Multiclass classifiers are important for handling practical classification problems that have more than two categories. Many of the multiclass classifiers are designed extending upon existing binary class classifiers utilizing one-versus-one, one-versus-rest, and the directed acyclic graph scheme [1]. For example, support vector machines (SVMs) [2,3] are originally developed for binary class problems and have been extended to handle multiclass problems.

Fisher's linear discriminant analysis (FDA) [4] was developed for dimension reduction of binary class problems and its extension to multiclass is generally referred to as Linear Discriminant Analysis (LDA). Unlike many ohter methods designed for multiclass problems, the LDA does not tackle a multiclass problem as a set multiple binary class problems. A limitation in the classical LDA is that it requires at least one scatter matrix be nonsingular. This condition will break down especially when the data set is undersampled, i.e., the data dimension is higher than the number of available data points. To overcome the nonsingularity restriction, LDA based on the generalized singular values decomposition (LDA/GSVD) has been introduced [5,6]. Several two-stage approaches [7–9] for dimension reduction have also been proposed to reduce computational complexity without the nonsingularity limitation of FDA. Recently, a comparison of generalized LDA algorithms has been studied [10]. Generalized LDA algorithms have also been used for dimension reduction and feature extraction [11,5,12] and have been extended with kernelized versions to deal with nonlinear problems [13–15] for multiclass problems.

For classification, methods such as the k-nearest neighbor (kNN) classifiers or SVMs have been in wide use. These methods require training and parameter optimization such as estimation of the k-value in the kNN classifier and estimation of the soft margin parameter in SVMs.

In this paper, we introduce nonparametric multiclass classifiers taking advantage of a dimension reducing transformation matrix obtained from generalized LDA. A two-stage generalized LDA algorithm is presented in Section 2, which require low computational complexity and low memory requirements for undersampled problems. The rest of the paper is organized as follows. In Section 3.1, multiclass centroid based linear and nonlinear discriminant classifiers based on generalized LDA, called CLDC/gLDA and CKDC/gLDA are introduced. In Section 3.2, we introduce a multiclass marginal linear discriminant classifier (MLDC) and derive its nonlinear version. We also introduce a Bayesian linear discriminant classifier (BLDC) and a one-dimensional BLDC (1BLDC) for multiclass classification in Section 3.3 and Section 3.4, respectively. In Section 4, we compare the test results of these classifiers in terms of accuracy and computational complexity.

The following notations are used throughout this paper: a given data set $A \in \mathbb{R}^{m \times n}$ with p classes is denoted as

$$A = \begin{pmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_m^T \end{pmatrix} = \begin{pmatrix} A_1 \\ \vdots \\ A_p \end{pmatrix} \in \mathbb{R}^{m \times n},$$

where $A_i \in \mathbb{R}^{m_i \times n}$ denotes the submatrix whose rows belong to class i. The jth row of A is denoted as $\mathbf{a}_j^T \in \mathbb{R}^{1 \times n}$. In addition, M_i denotes the set of row indices of the data items that belong to class i, m_i the number of items in class i, $\mathbf{c}_i \in \mathbb{R}^{n \times 1}$ a centroid vector which is the average of all the data in the class i, and $\mathbf{c}$ the global centroid vector. For any matrix A, A_{ik} denotes the (i,k)th element.

2 Linear Discriminant Analysis for Cluster Structure Preserving Dimension Reduction

The goal of linear discriminant analysis is to find a dimension reducing transformation that minimizes the scatter within each class and maximizes the scatter between classes in a reduced dimensional space. The within-class scatter matrix S_w, the between-class scatter matrix S_b, and the mixture scatter matrix S_m are defined as

$$S_w = \sum_{i=1}^{p} \sum_{j \in M_i} (\mathbf{a}_j - \mathbf{c}_i)(\mathbf{a}_j - \mathbf{c}_i)^T,$$

$$S_b = \sum_{i=1}^{p} m_i (\mathbf{c}_i - \mathbf{c})(\mathbf{c}_i - \mathbf{c})^T,$$

and

$$S_m = \sum_{i=1}^{m} (\mathbf{a}_i - \mathbf{c})(\mathbf{a}_i - \mathbf{c})^T.$$

Then we have

$$\mathrm{trace}(S_w) = \sum_{i=1}^{p} \sum_{j \in M_i} \|\mathbf{a}_j - \mathbf{c}_i\|_2^2$$

and

$$\mathrm{trace}(S_b) = \sum_{i=1}^{p} m_i \|\mathbf{c}_i - \mathbf{c}\|_2^2.$$

Defining the matrices

$$H_w = [A_1^T - \mathbf{c}_1 \mathbf{e}_1^T, \ldots, A_p^T - \mathbf{c}_p \mathbf{e}_p^T] \in \mathbb{R}^{n \times m}, \tag{1}$$

where $\mathbf{e}_i = [1, \ldots, 1]^T \in \mathbb{R}^{m_i \times 1}$,

$$H_b = [\sqrt{m_1}(\mathbf{c}_1 - \mathbf{c}), \ldots, \sqrt{m_p}(\mathbf{c}_p - \mathbf{c})] \in \mathbb{R}^{n \times p}, \tag{2}$$

and

$$H_m = [\mathbf{a}_1 - \mathbf{c}, \mathbf{a}_2 - \mathbf{c}, \ldots, \mathbf{a}_m - \mathbf{c}] \in \mathbb{R}^{n \times m}, \tag{3}$$

we have

$$S_w = H_w H_w^T, \;\; S_b = H_b H_b^T, \;\; \text{and} \;\; S_m = H_m H_m^T \tag{4}$$

When S_w is nonsingular, the simultaneous optimization, i.e. minimizing the scatter within each class and maximizing the scatter between classes, is commonly

approximated by maximizing

$$J_1(W) = \text{trace}((W^T S_w W)^{-1}(W^T S_b W)), \tag{5}$$

where $W \in \mathbb{R}^{n \times l}$ denotes the transformation that maps a vector in the n dimensional space to a vector in the l dimensional space. It is well know that $rank(S_w^{-1} S_b) \leq p - 1$ and columns of W that maximizes $J_1(W)$ of the leading $p - 1$ eigenvectors of $S_w^{-1} S_b$ [16].

For two-class problems, the dimension reducing transformation W is $n \times 1$, which we will denote as a vector $\mathbf{w}$,

$$\mathbf{w} = S_w^{-1}(\mathbf{c}_1 - \mathbf{c}_2).$$

However, when the number of features is larger than the number of examples ($n > m$), S_w is singular. LDA/GSVD [5] circumvents this nonsingularity restriction so that it can effectively reduce the dimension even when $n > m$.

We propose the following fast algorithm for finding a solution for the. For a similar approach, see [17]. Given a data matrix $A \in \mathbb{R}^{m \times n}$ with p classes, this algorithm computes the columns of the matrix $W \in \mathbb{R}^{n \times p}$, which preserves the class structure in the reduced dimensional space, and it also computes the p dimensional representation $Y \in \mathbb{R}^{m \times p}$ of A.

(1) Compute the SVD of H_m:

$$H_m = (U_s \, U_{n-s}) \begin{pmatrix} \Sigma_s & 0 \\ 0 & 0 \end{pmatrix} V^T = U D V^T$$

where $s = \text{rank}(H_m)$.

(2) Compute the QR decomposition of $\Sigma_s^{-1} U_s^T H_b$:

$$\Sigma_s^{-1} U_s^T H_b = Q_p R$$

Algorithm 1 LDA/EVD-QRD

Given a data matrix $A \in \mathbb{R}^{m \times n}$ ($m \ll n$) with p classes, this algorithm computes the columns of the matrix $W \in \mathbb{R}^{n \times p}$, which preserves the class structure in the reduced dimensional space, and it also computes the p dimensional representation $Y \in \mathbb{R}^{m \times p}$ of A.

(1) Compute $H_b \in \mathbb{R}^{n \times p}$ and $H_m \in \mathbb{R}^{n \times m}$ from A according to Eqns. 2 and 3, respectively.

(2) Compute the EVD of $H_m^T H_m \in \mathbb{R}^{m \times m}$:

$$H_m^T H_m = (\underbrace{V_1}_{s} \quad \underbrace{V_2}_{m-s}) \begin{pmatrix} \Sigma_H^2 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix},$$

where $s = \mathrm{rank}(H_m)$.

(3) Compute U_s from $H_m V_1 \Sigma_H^{-1}$.

(4) Compute the QR decomposition of $\Sigma_H^{-1} U_s^T H_b$: $\Sigma_H^{-1} U_s^T H_b = Q_p R$.

(5) Let $W = U_s \Sigma_H^{-1} Q_p$.

(6) $Y = AW$

(3) Let $W = U_s \Sigma_s^{-1} Q_p$.

(4) $Z = AW$

For computing the SVD of $H_m \in \mathbb{R}^{n \times m}$ when $m \ll n$, we first compute the reduced QRD of H_m, and then an SVD algorithm is applied to the upper triangular factor of size $m \times m$. The dimension reduction algorithm based on this procedure is referred as linear discriminant analysis based on QRD (LDA/QRD) in this paper. It is possible to reduce the computational complexity in LDA/QRD when $m \ll n$ by using the EVD of $H_m^T H_m \in \mathbb{R}^{m \times m}$ which can be represented as

$$H_m^T H_m = V D^T U^T U D V^T = V D^2 V^T.$$

Then, we have $U = H_m V D^{-1}$ and $\Sigma_s \in \mathbb{R}^{s \times s} = D_H(1{:}s, 1{:}s)$ where $s = \mathrm{rank}(H_m)$. This LDA/EVD-QRD algorithm is summarized in Algorithm 1.

3 Multiclass Discriminant Classifiers with Generalized LDA Algorithms

In this section, we present a way to take advantage of each column of the dimension reducing transformation matrix W for multiclass classifications. LDA/QRD or LDA/EVD-QRD produces a dimension reducing transformation $W \in \mathbb{R}^{n \times p}$ that has p columns. However, for the simple illustration of the multiclass classifiers, suppose that we obtained $W \in \mathbb{R}^{n \times (p-1)}$ by LDA/GSVD. We can obtain data points transformed from the original vector space to the reduced dimensional space by this W.

3.1 Multiclass Centroid Based Linear Discriminant Classifier (CLDC)

In the centroid based linear classification, a test data point, $\mathbf{x}$, is assigned to a class by finding the closest centroid. Using LDA/GSVD, we can efficiently obtain $W \in \mathbb{R}^{n \times (p-1)}$ without nonsingularity restriction. The $(p-1) \times 1$ centroid vector $\mathbf{c}_i^r$ for a class Ω_i in the reduced dimensional space is computed by

$$\mathbf{c}_i^r = \frac{1}{m_i} \sum_{j \in M_i} W^T \mathbf{a}_j,$$

where M_i is the set of data items that belong to the class Ω_i and m_i the number of data items in the class Ω_i. For a given sample data point, $\mathbf{x}$, the centroid based linear discriminant classifier with generalized LDA (CLDC/gLDA) assigns the class of $\mathbf{x}$ by

$$\arg \min_{1 \le i \le p} \|\mathbf{c}_i^r - W^T \mathbf{x}\|.$$

This CLDC has been widely used for multiclass classification based on LDA due to its simplicity [5,12].

Here, we also describe the centroid based kernel discriminant classifier with generalized LDA (CKDC/gLDA). Given a kernel function $K(\mathbf{x}, \mathbf{y})$, in CKDC, the cen-

troids of the training set in the reduced dimensional space are computed by

$$\mathbf{c}_i^r = \frac{1}{m_i} \sum_{j \in M_i} W^T K(A, \mathbf{a}_j),$$

where $W \in \mathbb{R}^{m \times (p-1)}$ is obtained from KDA [15] and $K(A, \mathbf{a}_j)$ is a $m \times 1$ kernel column vector and

$$K(A, \mathbf{a}_j) = \begin{bmatrix} K(\mathbf{a}_1^T, \mathbf{a}_j) \\ \vdots \\ K(\mathbf{a}_m^T, \mathbf{a}_j) \end{bmatrix}.$$

An example of one of the most commonly used kernel is the radial basis function (RBF) kernel $K(\mathbf{a}_i^T, \mathbf{a}_j) = \exp(-\gamma \|\mathbf{a}_i - \mathbf{a}_j\|^2)$.

For a given sample data point, $\mathbf{x}$, the class of the data point is assigned by

$$\arg \min_{1 \le i \le p} \|\mathbf{c}_i^r - W^T K(A, \mathbf{x})\|.$$

3.2 Multiclass Marginal Linear Discriminant Classifier (MLDC)

In this section, we first describe the marginal linear discriminant classifier for binary class problems and then show how it is extended to multiclass marginal classifiers. Suppose we have the dimension reducing transformation vector $\mathbf{w}$ from applying the LDA/GSVD to a problem with two classes. Then, we can use $\mathbf{w}$ to find a decision function of a separating hyperplane

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b,$$

so that $sign(f(\mathbf{x})) > 0$ when $\mathbf{x}$ belongs to the positive class and $sign(f(\mathbf{x})) \le 0$ when $\mathbf{x}$ belongs to the negative class, as in the SVM classifiers. The bias term b for this $f(x)$ based on LDA can be computed by $b = -\mathbf{w}^T \mathbf{c}$.

Now, we introduce a marginal modification of LDA where the bias term is com-

puted not based on the global centroid but based on cluster centroids. We define a negative class as that with a smaller centroid value $\mathbf{c}_-^r$ in the transformed one-dimensional space between the two classes, i.e. $\mathbf{c}_-^r < \mathbf{c}_+^r$. The two centroid vectors of the training set in the reduced dimensional space are

$$\mathbf{c}_-^r = \frac{1}{m_-} \sum_{i \in M_-} \mathbf{w}^T \mathbf{a}_i \quad \text{and} \quad \mathbf{c}_+^r = \frac{1}{m_+} \sum_{i \in M_+} \mathbf{w}^T \mathbf{a}_i,$$

where M_- and M_+ are the set of data items that belong to the negative class and the positive class respectively, and m_- and m_+ are the numbers of data points for each class respectively. The bias term is defined as the mean between a maximum value of the negative class and a minimum value of the positive class in the one-dimensional reduced space:

$$b = -\frac{1}{2}(\max_{i \in M_-}(\mathbf{w}^T \mathbf{a}_i) + \min_{i \in M_+}(\mathbf{w}^T \mathbf{a}_i)). \tag{6}$$

This method of choosing the bias b is the same as that of SVMs. The bias term is determined by the boundary points of each class. When a binary classification problem in the reduced dimensional space is completely separable, Eqn. (6) guarantees to find a class separating line, while $b = -\mathbf{w}^T \mathbf{c}$ may fail. In the above discussion, we proposed a method to determine the decision function $f(\mathbf{x}) sign(\mathbf{w}^T \mathbf{x} + b)$ for a binary class problem, where the weight vector $\mathbf{w}$ comes from generalized LDA that overcomes nonsingularity restriction and the bias b is calculated from Eqn. (6).

The above binary marginal classifier is now extended to a multiclass classifier. The dimension reducing transformation matrix W maps any data point in the original n dimensional space to the reduced $p-1$ dimensional space. Based on W, we form $t = p(p-1)/2$ binary decision functions for p classes. For example, when $p = 4$, all possible binary combinations of the classes are

$$\left\{ (\Omega_1, \Omega_2), (\Omega_1, \Omega_3), (\Omega_1, \Omega_4), (\Omega_2, \Omega_3), (\Omega_2, \Omega_4), (\Omega_3, \Omega_4) \right\} \tag{7}$$

where each (Ω_i, Ω_j) accounts for the binary decision function for classes i and j, and Ω_i denotes the ith class. The first binary decision function is related to the

decision between the class Ω_1 and Ω_2 among all $p(p-1)/2 = 6$ binary decision functions. In general, there are $p-1$ binary decision functions related to each class.

A bias value $b_k(\Omega_i, \Omega_j)$ of the binary decision function for class Ω_i and class Ω_j based on the kth column $\mathbf{w}_k$ of W is defined as

$$b_k(\Omega_i, \Omega_j) = -\frac{1}{2}(\max_{u \in M_\alpha}(\mathbf{w}_k^T \mathbf{a}_u) + \min_{v \in M_\beta}(\mathbf{w}_k^T \mathbf{a}_v)), \tag{8}$$

where M_α and M_β are the sets of data items that belong to the negative class (Ω_α) and the positive class (Ω_β) respectively. The class Ω_α is assumed to have the smaller mean value in the transformed one-dimensional space between the class Ω_i and the class Ω_j. The Ω_β is the other class. Then, a bias matrix $B \in \mathbb{R}^{t \times (p-1)}$ is formed as

$$B = \begin{pmatrix} b_1(1,2) & b_2(1,2) & \dots & b_{p-1}(1,2) \\ \vdots & \vdots & & \vdots \\ b_1(1,p) & b_2(1,p) & \dots & b_{p-1}(1,p) \\ b_1(2,3) & b_2(2,3) & \dots & b_{p-1}(2,3) \\ \vdots & \vdots & & \vdots \\ b_1(2,p) & b_2(2,p) & \dots & b_{p-1}(2,p) \\ \vdots & \vdots & & \vdots \\ b_1(p-1,p) & b_2(p-1,p) & \dots & b_{p-1}(p-1,p) \end{pmatrix}$$

where $b_k(\Omega_i, \Omega_j)$ is denoted as $b_k(i,j)$ for notational simplicity. Note that the index set in the first column would be what is listed in (7) when $p = 4$.

Corresponding negative class index Ω_α and positive class index Ω_β for each element of the matrix B are stored in the same location of a matrix $C^N \in \mathbb{R}^{t \times (p-1)}$ and a matrix $C^P \in \mathbb{R}^{t \times (p-1)}$, respectively.

Now, we present a method to classify a new test point $\mathbf{x}$ using W and B. We first define a decision matrix F as

$$F = sign(Y+B) = \begin{pmatrix} sign(\mathbf{w}_1^T\mathbf{x} + B_{1,1}) & \dots & sign(\mathbf{w}_{p-1}^T\mathbf{x} + B_{1,p-1}) \\ sign(\mathbf{w}_1^T\mathbf{x} + B_{2,1}) & \dots & sign(\mathbf{w}_{p-1}^T\mathbf{x} + B_{2,p-1}) \\ \vdots & & \vdots \\ sign(\mathbf{w}_1^T\mathbf{x} + B_{t,1}) & \dots & sign(\mathbf{w}_{p-1}^T\mathbf{x} + B_{t,p-1}) \end{pmatrix} \in \mathbf{R}^{t\times(p-1)},$$

where

$$Y = \begin{pmatrix} \mathbf{w}_1^T\mathbf{x},\ \mathbf{w}_2^T\mathbf{x},\ \dots,\ \mathbf{w}_{p-1}^T\mathbf{x} \\ \mathbf{w}_1^T\mathbf{x},\ \mathbf{w}_2^T\mathbf{x},\ \dots,\ \mathbf{w}_{p-1}^T\mathbf{x} \\ \vdots \quad \vdots \quad \vdots \\ \mathbf{w}_1^T\mathbf{x},\ \mathbf{w}_2^T\mathbf{x},\ \dots,\ \mathbf{w}_{p-1}^T\mathbf{x} \end{pmatrix}.$$

Then for each (i,k), if F_{ik} is negative, the class is assigned by C_{ik}^N, otherwise the class is assigned by C_{ik}^P. A decision vector $\mathbf{d}$ for the ith binary discriminant contains the $p-1$ class indexes obtained by the ith row vector in the matrix F. Then, the class assignment of the ith binary discriminant is determined as the dominant class in the vector $\mathbf{d}$. For example, let us assume that the first linear discriminant considers class Ω_1 and class Ω_2 among total 7 classes. If the first row vector of the matrix F is $F_{1\cdot} = [-1,1,0,1,1,-1]$, and corresponding row vectors of the class index matrices are $C_{1\cdot}^N = [1,2,1,1,2,2]$ and $C_{1\cdot}^P = [2,1,2,2,1,1]$, the decision vector $\mathbf{d}$ will be [1,1,0,2,1,2]. The class assignment of 0 means that the binary decision function cannot determine between two classes. The data point $\mathbf{x}$ is assigned to class Ω_1 since the class Ω_1 is dominant in the decision vector $\mathbf{d}$. When the ith binary discriminant is related with the classification between class Ω_1 and class Ω_2, the class assignment for the ith binary discriminant can be achieved by the following decision scheme:

NOTE: need to check the following carefull. val(C_{ik}^N), val(C_{ik}^P)??

(1) val(Ω_1)=0; val(Ω_2)=0;

(2) For k = 1 to $p-1$

- $\delta = sign(Y_{ik} + B_{ik})$;
- If $\delta < 0$ then val(C_{ik}^N) = val(C_{ik}^N) - δ

  else val(C_{ik}^P) = val(C_{ik}^P) + δ

(3) If val(Ω_1) $>$ val(Ω_2) then class(i) $\leftarrow \Omega_1$

else if val(Ω_1) $<$ val(Ω_2) then class(i) $\leftarrow \Omega_2$

else class(i) $\leftarrow 0$

If two classes can be completely separated, we chose the bias term by Eqn. (8). The transformed data points to the one-dimensional space in Eqn. (8) for the ith discriminant and the kth column of W can overlap when

$$\max_{u \in M_\alpha} (\mathbf{w}_k^T \mathbf{a}_u) > \min_{v \in M_\beta} (\mathbf{w}_k^T \mathbf{a}_v), \tag{9}$$

where M_α and M_β are the sets of data items that belong to the negative class (Ω_α) and the positive class (Ω_β), respectively. In this case, we use the following scheme to obtain a decision for a binary discriminant:

(1) δ = $\|W^T\mathbf{x} - \mathbf{c}_{\Omega_1}^r\|$ - $\|W^T\mathbf{x} - \mathbf{c}_{\Omega_2}^r\|$;

(2) If $\delta < 0$ then class(i) $\leftarrow \Omega_1$

else if $\delta > 0$ then class(i) $\leftarrow \Omega_2$

else class(i) $\leftarrow 0$

When class assignment 0, the decision function cannot discriminate between the two classes and this decision does not affect the voting for the final decision of multiclass classification. After repeating for all t binary discriminants, we obtain

$p(p-1)/2$ decisions. There are $p-1$ binary discriminants that are related to a class. One-against-one method was applied in order to assign a vector $\mathbf{x}$ to a class, which is a common way to extend binary classifiers such as SVMs to a multiclass classifier. Whenever we cannot decide one class due to the same number of decisions, we chose a lower index class.

Now, we introduce a marginal kernel discriminant classifier based on generalized LDA (MKDC/gLDA) using the kernel discriminant analysis (KDA) [15]. Now, we build $t = p(p-1)/2$ binary discriminants for each column of Y. A bias value $b_k(\Omega_i, \Omega_j)$ of the kernelized binary discriminant between class Ω_i and class Ω_j using the kth column of W is defined as

$$b_k(\Omega_i, \Omega_j) = -\frac{1}{2}(\max_{u \in M_\alpha}(\mathbf{w}_k^T K(A, \mathbf{a}_u)) + \min_{v \in M_\beta}(\mathbf{w}_k^T K(A, \mathbf{a}_v))),$$

where M_α and M_β are the set of data items that belong to the negative class (Ω_α) and the positive class (Ω_β) respectively, and $\mathbf{w}_k$ is the kth column of W. The class Ω_α has smaller mean value in the transformed one-dimensional space between the class Ω_i and the class Ω_j. The Ω_β is the other class. We build a bias matrix $B \in \mathbb{R}^{t \times (p-1)}$ and a transformed matrix $Y \in \mathbb{R}^{t \times (p-1)}$ whose element can be computed by

$$Y_{ik} = \mathbf{w}_k^T K(A, \mathbf{x}). \tag{10}$$

Then, we assign class of a new test data point $\mathbf{x}$ by the similar way described in MLDC.

3.3 Multiclass Bayesian Linear Discriminant Classifier (BLDC)

The conditional probabilities $P(\Omega_i|\mathbf{x})$ can be obtained by

$$P(\Omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\Omega_i)P(\Omega_i)}{P(\mathbf{x})} = \frac{\text{distribution probability} \cdot \text{prior}}{\text{evidence}}.$$

If we have only two classes, then we can define a discriminant function as:

$$d_2(\mathbf{x}) = \log\left(\frac{P(\Omega_2|\mathbf{x})}{P(\Omega_1|\mathbf{x})}\right),$$

where the class membership is determined based on $sign(d_2(x))$. Now suppose that we have two classes Ω_1 and Ω_2, and the feature value is normally distributed within each class. That is:

$$P(\mathbf{x}|\Omega_1) = \frac{1}{\sqrt{(2\pi)^d \cdot |\Sigma_1|}} \exp\left(\frac{(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)}{-2}\right),$$

where μ_1 is the mean vector in the class Ω_1, Σ_1 is the corresponding covariance matrix, and $|\Sigma_1|$ denotes the determinant of Σ_1. If the other class Ω_2 is also normally distributed, The discriminant function will be

$$\begin{aligned} d_2(\mathbf{x}) = &\frac{1}{2}\mathbf{x}^T(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{x} + (\Sigma_2^{-1}\mu_2 - \Sigma_1^{-1}\mu_1)^T\mathbf{x} \\ &+\frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2) \\ &+\frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|} - \log\frac{P(\Omega_1)}{P(\Omega_2)}. \end{aligned}$$

Assuming $\Sigma_1 = \Sigma_2 = \Sigma$ and the log of the class likelihood, i.e. $\log(P(\Omega_1)/P(\Omega_2))$, is 0, we can simplify the line discriminant function to

$$d_2(\mathbf{x}) = (\mu_2 - \mu_1)^T\Sigma^{-1}\left(\mathbf{x} - \frac{\mu_1 + \mu_2}{2}\right). \tag{11}$$

After obtaining the dimension reducing transformation W by generalized LDA, the $p \times 1$ mean vector μ_i of a class Ω_i in the reduced dimensional space can be computed by

$$\mu_i = \frac{1}{m_i}\sum_{j \in M_i} W^T\mathbf{a}_j,$$

where M_i is the set of row indices of the data set A that belong to the class Ω_i and m_i the number of data points in the class Ω_i. After repeating for all $t = p(p-1)/2$ linear discriminants for p classes problem, we obtain $p(p-1)/2$ decisions. There

are $p-1$ binary discriminants that are related to a class. By voting, a vector $\mathbf{x}$ is assigned to the most frequent class.

3.4 Multiclass One-dimensional Bayesian Linear Discriminant Classifier (1BLDC)

Here is another approach to divide a decision for a binary discriminant in the p-dimensional reduced space into p one-dimensional decisions again like MLDC. To discriminate this method with BLDC described in the previous section, we call this approach one-dimensional BLDC (1BLDC). For the ith discriminant and kth column of W, the mean value of the class Ω_j can be computed by

$$\mu_j = \frac{1}{m_j} \sum_{u \in M_j} \mathbf{w}_k^T \mathbf{a}_u,$$

where M_j is the set of data items that belong to the class Ω_j and m_j is the number of data points in the class Ω_j. Assuming the one-dimensional feature is normally distributed with each class,

$$P(x|\Omega_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(\frac{(x-\mu_j)^2}{-2\sigma_j^2}\right),$$

where σ_j is the standard deviation, the discriminant function between class Ω_1 and class Ω_2 is

$$\begin{aligned} d_2(x) = & \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) x^2 + \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right) x \\ & + \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) + \frac{1}{2}\log\frac{\sigma_1^2}{\sigma_2^2} - \log\frac{P(\Omega_1)}{P(\Omega_2)} \end{aligned}$$

In between the two roots, d_2 is negative, indicating the choice of the class Ω_1. Beyond either root, d_2 is positive, indicating the choice of the class Ω_2. Assuming $\sigma_1 = \sigma_2 = \sigma$ and the log of the class likelihood, i.e. $\log(P(\Omega_1)/P(\Omega_2))$, is 0, we can simplify the linear discriminant function to

$$d_2(x) = \frac{(\mu_2 - \mu_1)}{\sigma^2}\left(x - \frac{\mu_1 + \mu_2}{2}\right). \tag{12}$$

Table 1

Ten-fold cross-validation (CV) testing accuracy (%) and the number of data points, features, and extracted features by different dimension reduction algorithms on AMLALL data set of 7129 genes and 72 samples.

Methods	10-fold CV	data	features	features
	EVD-QRD/QRD (GSVD)		(used)	(extracted)
gLDA & SVMs	95.89% / 95.89% (95.89%)	72	7129	2/2 (1)
gLDA & kNN	95.89% / 95.89% (95.89%)	72	7129	2/2 (1)
MLDC	95.89% / **98.57%** (95.89%)	72	7129	2/2 (1)
CLDC	95.89% / 95.89% (95.89%)	72	7129	2/2 (1)
1BLDC	95.89% / 95.89% (95.89%)	72	7129	2/2 (1)
BLDC	95.89% / 95.89% (95.89%)	72	7129	2/2 (1)

This linear discriminant classifier is applied for each reduced feature in one-dimensional space at a time, i.e. $d_2(\mathbf{w}_k^T\mathbf{x})$ for $1 \leq k \leq p-1$. The class assignment for the ith binary discriminant can be performed by the same method used for MLDC with $\delta = d_2(\mathbf{w}_k^T\mathbf{x})$. After repeating for all $t = p(p-1)/2$ binary discriminants for p classes problem, we obtain $p(p-1)/2$ decisions. By voting, a vector $\mathbf{x}$ is assigned to the most frequent class.

4 Results and Discussion

In this section, we present experimental results for the purpose of comparing various multiclass classification methods presented in this paper to those existing methods such as SVM and kNN when they are combined with dimension reduction. The test results are generated based on five different data sets. The first is the leukemia

data set (AMLALL) [18] that consists of 7129 genes and 72 samples, which is split into a training set of 38 samples (27 samples are acute lymphoblastic leukemia (ALL) and 11 samples are acute myeloid leukemia (AML)), and a test set of 34 samples (20 belong to ALL and 14 belong to AML). The Ionosphere, Wine, and ISOLET (Isolated Letter Speech Recognition) data sets are from UC Irvine test problems [19]. Since Ionosphere data set, which has a classification of radar returns from the ionosphere, is well known to be hardly linearly separable, they have often been used for estimating performances of the non-linear classifiers. The wine data set determines the origin of wines using chemical analysis. The ISOLET1559 data set that has 1559 instances as well as the full ISOLET data set of 7797 instances were used for ten-fold cross validation. We performed ten-fold cross validation test to compare several classification algorithms.

In Table 1, we presented ten-fold cross validation results on AMLALL data set when the number of features is larger than the number of data points. We refer the methods of linear feature extraction by generalized LDA and classification in the reduced feature space by SVMs and kNN classifier to gLDA & SVMs and gLDA & kNN. (NOTE: why SVMs and gLDA&SVMs? SVMs mean full space SVMs? no such resutls presented) The regularization parameters for SVMs and gLDA & SVMs were set to $C = 2^{-4}$ and $C = 2^4$ respectively. (NOTE: why RR here? never mentioned it before, what is λ? undefined w.r.t. RR) For ridge regression (RR) [20,21], we used the regularization parameter of $\lambda = 1/2^{-4}$. For k-nearest neighbor classification, we set the number of neighbors to 15 for class assignment of test data points. The LDA/EVD-QRD and LDA/QRD are variations of the two-stage approach with PCA and LDA. Even though LDA/EVD-QRD and LDA/QRD are algorithmically equivalent, they produced different results when we use the marginal classifier. This is due to different numerical behavior and possibility of information loss in forming $H_m^T H_m$. There was no significant difference of ten-fold cross validation accuracy for the six classifiers and different three dimension reduction algorithms, i.e. LDA/EVD-QRD, LDA/QRD, and LDA/GSVD.

Table 2

Ten-fold cross-validation (CV) testing accuracy (%) on the binary and multiclass data sets by linear classifiers. All values are given by using LDA/GSVD dimension reduction algorithm.

Methods	Ionosphere	Wine	ISOLET1559	ISOLET
	2 classes	3 classes	26 classes	26 classes
	(351 × 34)	(178 × 13)	(1559 × 617)	(7797 × 617)
gLDA & kNN	86.89%	98.86%	87.55%	93.78%
MLDC	**88.89%**	**99.44%**	**95.44%**	93.22%
CLDC	87.17%	98.86%	87.17%	92.00%
1BLDC	87.17%	**99.44%**	91.47%	**94.19%**
BLDC	87.17%	**99.44%**	85.82%	94.05%

From Table 1 and Table 2, we observed that our proposed multiclass classifiers could produce accuracy results comparable to that of kNN classifier. For ISOLET1559 data set, the marginal linear discriminant classifier produced highly accurate results of accuracy above 95% while other classifiers did not perform well. The 1BLDC performs similar or better than BLDC for all data sets. However, the MLDC performed better than the 1BLDC for some data sets and we believe it is due to its utilization of the marginal approach.

Table 2 shows that the proposed multiclass classifiers produce results as good as those using kNN classifiers. Even though there was no general winner across all data sets in terms of accuracy, MLDC and 1BLDC show better result than CLDC in general. The results suggest that building multiclass classifiers utilizing the dimension reducing transformation matrix is potentially another succssful approach for accurate nonparametric multiclass classification.

5 Conclusion

We proposed multi-class classification methods where the dimension reducing transformation matrix is directly utilized for multiclass classification. This procedure is computationally more efficient than using external k-nearest neighbor classification or support vector machines in the reduced dimensional space, which need to optimize parameters, such as the k-value in kNN classifier and the regularization/kernel parameters in SVMs. We proposed several multiclass classifiers based on generalized LDA. The multiclass marginal linear discriminant classifier and Bayesian linear discriminant classifier showed the comparable results to those of kNN classification or centroid-based classification followed by dimension reduction. The proposed multiclass classifiers can utilize any other two-stage approaches [7–9] of the generalized LDA algorithms.

Acknowledgement

The authors would like to thank the University of Minnesota Supercomputing Institute (MSI) for providing the computing facilities. The work of Haesun Park has been performed while at the National Science Foundation (NSF) and was partly supported by IR/D from the NSF. This material is based upon work supported in part by the National Science Foundation Grants CCR-0204109 and ACI-0305543. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] J. C. Platt, N. Cristianini, J. Shawe-Taylor, Large margin DAGs for multiclass classification, in: Advances in Neural Information Processing Systems, Vol. 12, MIT Press, 2000, pp. 547–553.

[2] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[3] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, 1998.

[4] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7, Part II (1936) 179–188.

[5] P. Howland, M. Jeon, H. Park, Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition, SIAM J. Matrix Anal. Appl. 25 (1) (2003) 165–179.

[6] P. Howland, H. Park, Generalizing discriminant analysis using the generalized singular value decomposition, IEEE Trans. Pattern Anal. Machine Intell. 26 (8) (2004) 995–1006.

[7] L. Chen, H. M. Liao, M. Ko, J. Lin, G. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (2000) 1713–1726.

[8] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data with application to face recognition, Pattern Recognition 34 (2001) 2067–2070.

[9] J. Yang, J. Y. Yang, Why can LDA be performed in PCA transformed space?, Pattern Recognition 36 (2003) 563–566.

[10] C. H. Park, H. Park, A comparison of generalized LDA algorithms for undersampled problems, Tech. Rep. 03-048, Department of Computer Science and Engineering, University of Minnesota (2003).

[11] J. Ma, S. Perkins, J. Theiler, S. Ahalt, Modified kernel-based nonlinear feature extraction, in: International Conference on Machine Learning and Application (ICMLKA'02), Las Vegas, NV, USA, 2002.

[12] H. Kim, P. Howland, H. Park, Dimension reduction in text classification using support vector machines, *Journal of Machine Learning Research*, to appear (2004).

[13] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K. R. Müller, Fisher discriminant analysis with kernels, in: Y. H. Hu, J. Larsen, E. Wilson, S. Douglas (Eds.), Neural Networks for Signal Processing IX, IEEE, 1999, pp. 41–48.

[14] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Computation 12 (10) (2000) 2385–2404.

[15] C. H. Park, H. Park, Kernel discriminant analysis based on the generalized singular value decomposition, Tech. Rep. 03-017, Department of Computer Science and Engineering, University of Minnesota (2003).

[16] K. Fukunaga, Introduction to Statistical Pattern Recognition, second edition, Academic Press, Boston, 1990.

[17] C. H. Park, H. Park, A fast dimension reduction algorithm with applications on face recognition and text classification, Tech. Rep. 03-050, Department of Computer Science and Engineering, University of Minnesota (2003).

[18] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

[19] C. L. Blake, C. J. Merz, UCI repository of machine learning databases (1998).
URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`

[20] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.

[21] N. Cristianini, J. Shawe-Taylor, Support Vector Machines and other kernel-based learning methods, University Press, Cambridge, 2000.